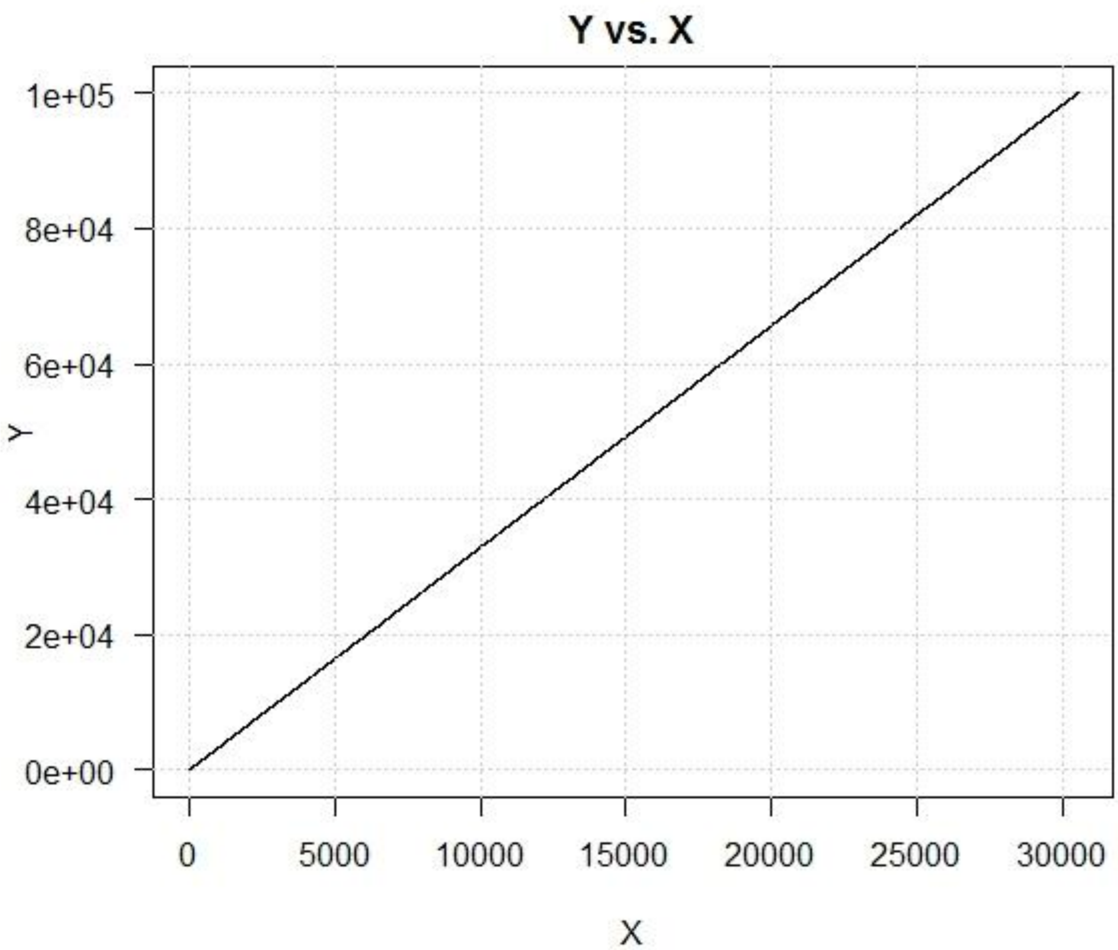
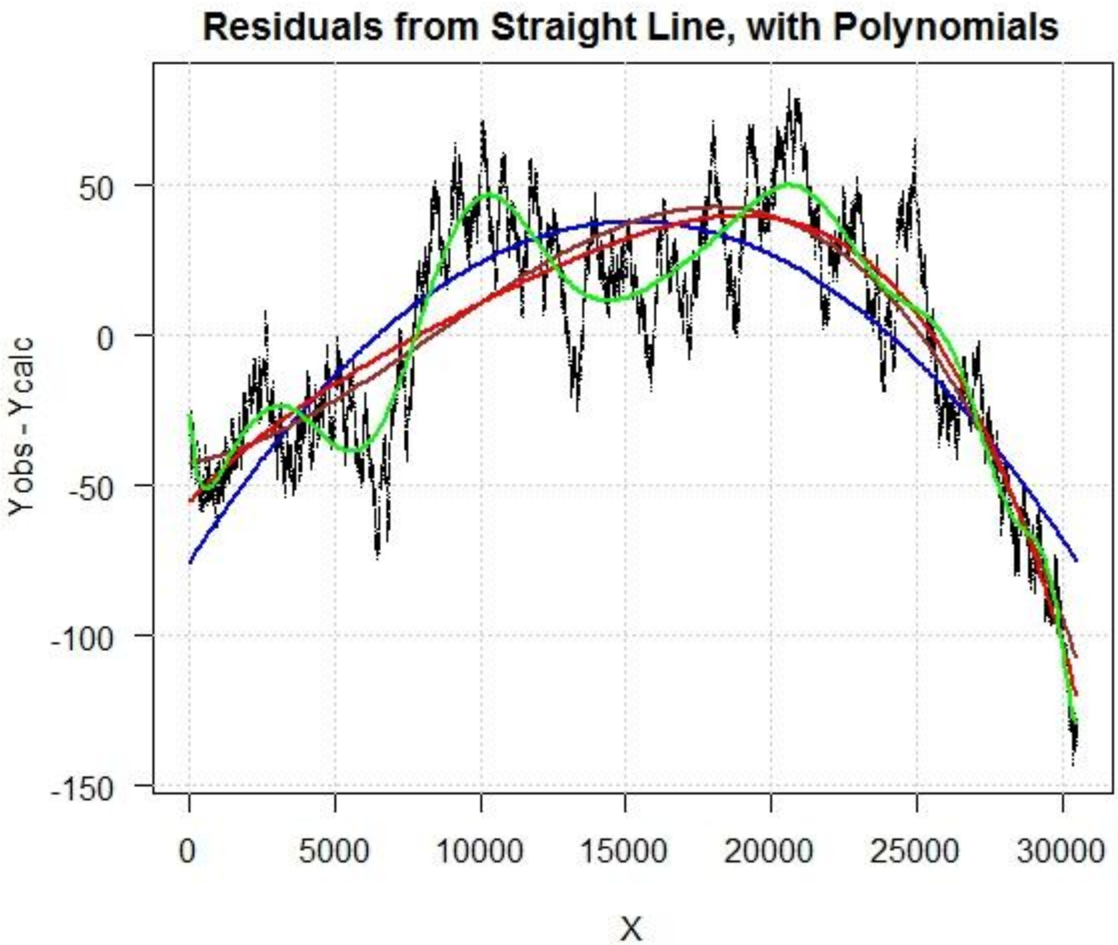
Dear Alex,

That’s an interesting series, and I’m curious to know what mathematical formula or algorithm generates it.

The first 30542 entries in the series certainly seem to be well represented by a straight line, with an intercept near zero, and a slope of about 3.3, as seen in Figure 1:



It gets very interesting when you look at the departures (Yobs – Ycalc) from the straight line, as seen in Figure 2, along with the fitted parabola (blue), cubic (brown) quartic (red), and 20^{th}-order polynomials:



Most of the observed Y values fall within +/- 100 of the fitted straight line. The departures seem to resemble some kind of fractal random walk, with no obvious pattern. So it's unlikely that fitting polynomials of any reasonable order will give you a formula for predicting Y values to, say, the +/- 1 or 2 units.

I don't know if the downward trend for X > 25000 is real, or just random fluctuation, with the values eventually coming back toward the straight line for X>30542. It would be interesting to see what happens for X = 40000 or 50000. Perhaps you could generate a few elements of this series for very large X (perhaps using Mathematica), to see how the series behaves beyond X = 30542.

Here are the first 20 order polynomials fitted to all 30542 data points in the file. To avoid getting extremely small coefficients in the formulas, I redefined X as the original X divided by 10000. So the coefficient of X in these formulas should be divided by 10000; the coefficient of X^2 should be divided by 10000^2; etc., to get the coefficients in terms of the original X variable.

I'm displaying the regression coefficients only to a few digits of precision; I could provide them to 15 digit precision if you need them.

Y = 29.5 + 32775*X
AIC=314703.5, BIC=314728.5, R^2=0.9999979, Adj.R^2=0.9999979, RMS Err=41.80328

Y = -46 + 32923*X - 48.6*X^2
AIC=282333.8, BIC=282367.1, R^2=0.9999993, Adj.R^2=0.9999993, RMS Err=24.60716

Y = -13.1 + 32794*X + 57.2*X^2 - 23.1*X^3
AIC=273329.4, BIC=273371, R^2=0.9999995, Adj.R^2=0.9999995, RMS Err=21.23415

Y = -26 + 32878*X - 66.7*X^2 + 40*X^3 - 10.3*X^4
AIC=272064.1, BIC=272114, R^2=0.9999995, Adj.R^2=0.9999995, RMS Err=20.79849

Y = -11.6 + 32737*X + 257*X^2 - 243*X^3 + 93.8*X^4 - 13.6*X^5
AIC=270709.2, BIC=270767.4, R^2=0.9999995, Adj.R^2=0.9999995, RMS Err=20.3419

Y = 5.12 + 32507*X + 1009*X^2 - 1228*X^3 + 699*X^4 - 188*X^5 + 19*X^6
AIC=269084.1, BIC=269150.7, R^2=0.9999995, Adj.R^2=0.9999995, RMS Err=19.80755

Y = -9.7 + 32779*X - 191*X^2 + 956*X^3 - 1267*X^4 + 739*X^5 - 200*X^6 + 20.5*X^7
AIC=267926, BIC=268000.9, R^2=0.9999995, Adj.R^2=0.9999995, RMS Err=19.43522

Y = -44.5 + 33599*X - 4890*X^2 + 12237*X^3 - 15118*X^4 + 10171*X^5 - 3803*X^6 + 743*X^7 - 59.1*X^8
AIC=261553.4, BIC=261636.7, R^2=0.9999996, Adj.R^2=0.9999996, RMS Err=17.50954

Y = -41 + 33496*X - 4149*X^2 + 9973*X^3 - 11504*X^4 + 6858*X^5 - 1995*X^6 + 163*X^7 + 41.7*X^8 - 7.34*X^9
AIC=261491.5, BIC=261583.1, R^2=0.9999996, Adj.R^2=0.9999996, RMS Err=17.49152

Y = -11.2 + 32421*X + 5350*X^2 - 25963*X^3 + 60555*X^4 - 78071*X^5 + 59796*X^6 - 27912*X^7 + 7797*X^8 - 1199*X^9 + 78*X^10
AIC=256942.4, BIC=257042.3, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=16.23594

Y = 12 + 31421*X + 15997*X^2 - 74755*X^3 + 180355*X^4 - 253783*X^5 + 222789*X^6 - 125928*X^7 + 45905*X^8 - 10441*X^9 + 1349*X^10 - 75.7*X^11
AIC=254121, BIC=254229.3, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.50282

Y = 3.92 + 31834*X + 10785*X^2 - 46319*X^3 + 96573*X^4 - 104570*X^5 + 51810*X^6 + 4307*X^7 - 20719*X^8 + 12180*X^9 - 3539*X^10 + 533*X^11 - 33.2*X^12
AIC=253788.5, BIC=253905.1, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.4184

Y = 3.92 + 31834*X + 10785*X^2 - 46319*X^3 + 96573*X^4 - 104570*X^5 + 51810*X^6 + 4307*X^7 - 20719*X^8 + 12180*X^9 - 3539*X^10 + 533*X^11 - 33.2*X^12 + ???*X^13
AIC=253788.5, BIC=253905.1, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.4184

Y = 1.49 + 31978*X + 8679*X^2 - 32950*X^3 + 50540*X^4 - 8119*X^5 - 79431*X^6 + 124567*X^7 - 96081*X^8 + 44347*X^9 - 12623*X^10 + 2106*X^11 - 167*X^12 + ???*X^13 + 0.609*X^14
AIC=253760.2, BIC=253885.1, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.41101

Y = 1.49 + 31978*X + 8679*X^2 - 32950*X^3 + 50540*X^4 - 8119*X^5 - 79431*X^6 + 124567*X^7 - 96081*X^8 + 44347*X^9 - 12623*X^10 + 2106*X^11 - 167*X^12 + ???*X^13 + 0.609*X^14 + ???*X^15
AIC=253760.2, BIC=253885.1, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.41101

Y = -5.74 + 32464*X + 586*X^2 + 25627*X^3 - 180115*X^4 + 546830*X^5 - 951595*X^6 + 1054828*X^7 - 781717*X^8 + 393349*X^9 - 132392*X^10 + 27990*X^11 - 3032*X^12 + ???*X^13 + 28.6*X^14 + ???*X^15 - 0.271*X^16
AIC=253494.3, BIC=253627.5, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.34382

Y = -5.74 + 32464*X + 586*X^2 + 25627*X^3 - 180115*X^4 + 546830*X^5 - 951595*X^6 + 1054828*X^7 - 781717*X^8 + 393349*X^9 - 132392*X^10 + 27990*X^11 - 3032*X^12 + ???*X^13 + 28.6*X^14 + ???*X^15 - 0.271*X^16 + ???*X^17
AIC=253494.3, BIC=253627.5, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.34382

Y = -9.53 + 32750*X - 4748*X^2 + 68983*X^3 - 372165*X^4 + 1068132*X^5 - 1879484*X^6 + 2181434*X^7 - 1733154*X^8 + 952980*X^9 - 356759*X^10 + 85460*X^11 - 10718*X^12 + ???*X^13 + 149*X^14 + ???*X^15 - 2.75*X^16 + ???*X^17 + 0.0298*X^18
AIC=253422.2, BIC=253563.8, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.32548

Y = -9.53 + 32750*X - 4748*X^2 + 68983*X^3 - 372165*X^4 + 1068132*X^5 - 1879484*X^6 + 2181434*X^7 - 1733154*X^8 + 952980*X^9 - 356759*X^10 + 85460*X^11 - 10718*X^12 + ???*X^13 + 149*X^14 + ???*X^15 - 2.75*X^16 + ???*X^17 + 0.0298*X^18 + ???*X^19
AIC=253422.2, BIC=253563.8, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.32548

Y = 3.43 + 31667*X + 17650*X^2 - 132927*X^3 + 621309*X^4 - 1933353*X^5 + 4083052*X^6 - 5926717*X^7 + 5970198*X^8 - 4173069*X^9 + 1984621*X^10 - 603874*X^11 + 96450*X^12 + ???*X^13 - 2232*X^14 + ???*X^15 + 75.4*X^16 + ???*X^17 - 1.96*X^18 + ???*X^19 + 0.0258*X^20
AIC=252545.7, BIC=252695.6, R^2=0.9999997, Adj.R^2=0.9999997, RMS Err=15.10688

Every additional term added to the polynomial increases its ability to fit the wiggles in this data (according to AIC and BIC), but the RMS Error term (indicating the root-mean-square average discrepancy between Yobs and Ycalc) doesn't come down by very much.

The main point is that it's very difficult to remove the pseudo-random variability in the data points by adding more terms to the regression model. I fitted a 100-order polynomial to the data, with the following results:

AIC=244964.5, BIC=245205.9, R^2=0.9999998, Adj.R^2=0.9999998, RMS Err=13.34123

While the AIC and BIC clearly indicate that a 100-order polynomial fits better than a 20-order polynomial, the improvement is not very impressive -- the RMS Error was improved by less than two units, only down to 13.3 units.

I'm sure that with a sufficiently high-order polynomial, you could match all the wiggles in Figure 2 with a much smaller RMS Error, but this same function would probably be totally unimpressive for even larger datasets. High-order polynomials cannot safely be used to extrapolate beyond the range of data from which they were generated.

I'm not sure how to interpret the ??? coefficients that appear in some of the terms of some polynomials of order 13 and above. The program gave "NA" values for these coefficients, but was still able to calculate other things about the regression, like AIC, BIC, R^2, etc. When I fitted the data using orthogonal polynomials instead of raw polynomials, this problem didn't occur.